DOCUMENT RESUME

ED 146 014 SE 022 861

AUTHOR Newman, Isadore; Fraas, John W.
TITLE Some Applied Research Concerns Using Multiple Linear Regression Analysis.
PUB DATE Apr 77
NOTE 51p.; Paper presented at the Ohio Academy of Science, Psychology Division, Capital University, Columbus, Ohio

EDRS PRICE MF-$0.83 HC-$3.50 Plus Postage.
DESCRIPTORS *Multiple Regression Analysis; Research Methodology; *Research Tools; *Social Sciences; *Statistical Analysis; Statistics

ABSTRACT

The intention of this paper is to provide an overall reference on how a researcher can apply multiple linear regression in order to utilize the advantages that it has to offer. The advantages and some concerns expressed about the technique are examined. A number of practical ways by which researchers can deal with such concerns as correlation/causation, upward bias R-squared, and multicollinearity are discussed. (MS)

****
* Documents acquired by ERIC include many informal unpublished *
* materials not available from other sources. ERIC makes every effort *
* to obtain the best copy available. Nevertheless, items of marginal *
* reproducibility are often encountered and this affects the quality *
* of the microfiche and hardcopy reproductions ERIC makes available *
* via the ERIC Document Reproduction Service (EDRS). EDRS is not *
* responsible for the quality of the original document. Reproductions *
* supplied by EDRS are the best that can be made from the original. *
****

ED146014

U.S. DEPARTMENT OF HEALTH, EDUCATION & WELFARE
NATIONAL INSTITUTE OF EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

SOME APPLIED RESEARCH CONCERNS USING MULTIPLE LINEAR REGRESSION ANALYSIS

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

Isadore Newman

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC) AND USERS OF THE ERIC SYSTEM."

Isadore Newman
The University of Akron

John W. Fraas
Ashland College

This paper was presented at The Ohio Academy of Science, Psychology Division; Capital University, Columbus, Ohio, April 1977.

SE 022 861

SOME APPLIED RESEARCH CONCERNS
USING MULTIPLE LINEAR REGRESSION ANALYSIS

Introduction

During the last fifteen years, multiple linear regression, the general case of the least squares solution, has developed into a dominate research technique for the social sciences. With this increase in usage of multiple linear regression, there have developed two opposing viewpoints with regard to its usefulness and its appropriateness. The arguments of both groups, the advocates and the critics of multiple linear regression, can be found in the recent literature. The advocates of multiple linear regression state and defend the advantages provided to the researcher who uses multiple linear regression. The critics state a variety of limitations and concerns with respect to utilizing multiple linear regression as a research technique.

The purpose of this paper is to examine the advantages claimed by the advocates of multiple linear regression and some of the concerns expressed by its critics. More than anything else what the authors of this paper have attempted to provide is an overall reference on how a researcher can apply multiple linear regression in order to utilize

the advantages that it has to offer. Also, the authors have attempted to provide a number of meaningful and practical methods by which researchers can deal with the concerns that are often cited by the critics of multiple linear regression, which are correlation/causation, upward bias R^2, and multicollinearity.

Advantages of Using Multiple Linear Regression

While a great deal of money and time is currently being directed toward research, there appears to be a general lack of acceptance of the relevance of research findings. One reason for the present skepticism has been that the statistical models used by researchers have frequently been unrelated or tangentially related to the research question of interest. There are a variety of reasons for this lack of agreement between the research question of interest and the statistical model.

One such reason is that courses that teach research methods generally emphasize data analysis, rather than practicing appropriate methods and procedures for asking and developing research questions. These courses do not adequately develop the skills of evaluating the research question and the statistical models that are most capable of reflecting the research question.

Quite often, a student coming out of these courses tends to select a familiar, "canned" standard statistical design, or package (cookbook approach) such as a 2 x 3, or

2 x 2 x 3, because he has not been taught to develop his own models to reflect their research question. Therefore, he uses these standard models which dictate the question being investigated. Sometimes a researcher is aware that these models do not completely represent his true research question. In addition, a significant F-value on a factorial design is often difficult to interpret. When this happens, he may then make inferential jumps from his data. These inferences may well be inappropriate.

Therefore, in many cases the researcher is unaware that his models are not really reflective of his research questions; and quite often, the unsophisticated researcher allows the statistical model to totally dictate his research question. Under these conditions, we find research that is technically correct but is not relevant because it is not related in a pragmatic way to a specific problem. (Newman, et al., 1976)

1. One advantage of using regression procedures is that the researcher finds it necessary to first state his hypothesis and then write the regression model needed to test that hypothesis. Thus, every test of significance is directly testing a specific question posed by the researcher. Also, regression is more flexible in allowing the researcher to write the models that specifically reflect his question of interest. The advantages provided by this flexibility can be seen in research questions that deal with interaction variables, directional and partial

interaction covariance, trend analysis, and questions that encounter the problem of disproportionality.

2. In dealing with interaction variables, a researcher with regression can ask interaction questions between catagorical variables, between catagorical and continuous variables or between continuous variables. Since regression can deal with catagorical and continuous variables, it is more flexible in its ability to reflect real-world problems than other statistical procedures. With regression, there is no need to catagorize variables that are continuous in nature as required, for example, by traditional ANOVA; therefore, one would not lose degrees of freedom and power. (McNeil, Kelly, McNeil, 1975, Kerlinger, 1973)).

An example of how a hypothesis which involves the group membership could be tested is listed below:

Example 1:

Y_1 = posttest score

X_1 = control group

X_2 = experimental group

X_3 = I.Q. score

X_4 = $X_1 * X_3$ (I.Q. scores for the students in the control group)

X_5 = $X_2 * X_3$ (I.Q. scores for the students in the experimental group)

$E_{1,2}$ = the error for each subject

U = the unit vector

$a_o, \ldots, a_y$ = partial regression coefficients

R_F^2 = variance in Y_1 accounted for by the full model

R_R^2 = variance in Y_1 accounted for by the restricted model

df_n = the number of linearly independent vectors in the full model minus the number of linearly dependent vectors in the restricted model.

df_d = the number of subjects minus the number of linearly independent vectors in the full model.

H_1 = the differences between the posttest scores of the control group and the experimental group are not constant across the range of I.Q. scores.

Model 1

$$Y_1 = a_0U + a_1X_1 + a_2X_2 + a_4X_4 + a_5X_5 + E_1$$

Restrictions: $a_4 = a_5 = a_3$

Model 2 $\quad Y_1 = a_0U + a_1X_1 + a_2X_2 + a_3X_3 + E_2$

By testing Model 1 against Model 2, that is, by determining if the F-value calculated by:

$$F = \frac{\left(R_F^2 - R_R^2\right) / df_n}{\left(1 - R_F^2\right) / df_d}$$

is significant, the researcher could determine if there is a significant interaction between the continuous variable of I.Q. scores and the categorical variables of the groups (McNeil, et al., 1975).

3. Regression also allows the researcher to test directional and partial interaction questions (McNeil, at al., 1975). For example, the researcher may hypothesize that I.Q. scores have a greater impact on the

posttest scores of the subjects in the experimental group than it does for the control group subjects. The researcher could obtain the answer to his research question by testing Model 1 against Model 2 (using the same variables and models listed previously). If a significant F-value was obtained and if $a_5 > a_4$, the research could conclude that I.Q. scores had a greater impact on posttest scores for the subjects of experimental group than for the subjects of the control group.

Regression also allows the research to test interaction questions that the researcher would tend not to ask if he was not familiar with regression procedures, that is, partial interaction questions (McNeil, et al., 1975). For example, a researcher might be interested in testing the following hypothesis (Fraas, 1977):

Example 2:

H_1 = Previous economic training has a greater impact on the average posttest scores of the students in the two experimental groups than for students in the two control groups. (Note: More than two groups could be used.)

Y_1 = posttest scores

X_1 = previous economic training (1 if yes; 0 otherwise)

X_2 = no previous economic training (1 if yes; 0 otherwise)

X_3 = Control Group I (1 if yes; 0 otherwise)

X_4 = Experimental Group I (1 if yes; 0 otherwise)

X_5 = Control Group II (1 if yes; 0 otherwise)

X_6 = Experimental Group II (1 if yes; 0 otherwise)

$X_7 = X_1 * X_3$ Students in Control Group I with previous economic training (1 if yes; 0 otherwise)

$X_8 = X_1 * X_4$ Students in Experimental Group I with previous economic training (1 if yes; 0 otherwise)

$X_9 = X_1 * X_5$ Students in Control Group II with previous economic training (1 if yes; 0 otherwise)

$X_{10} = X_1 * X_6$ Students in Experimental Group II with previous economic training (1 if yes; 0 otherwise)

$X_{11} = X_2 * X_3$ Students in Control Group I with no previous economic training (1 if yes; 0 otherwise)

$X_{12} = X_2 * X_4$ Student in Experimental Group I with no previous economic training (1 if yes; 0 otherwise)

$X_{13} = X_2 * X_5$ Students in Control Group II with no previous economic training (1 if yes; 0 otherwise)

$X_{14} = X_2 * X_6$ Students in Experimental Group II with no previous economic training (1 if yes; 0 otherwise)

$X_{15} = X_8 + X_7$

$X_{16} = X_9 - X_7$

$X_{17} = X_{10} + X_7$

$X_{18} = X_{11} + X_7$

$X_{19} = X_{12} - X_7$

$X_{20} = X_{13} + X_7$

$X_{21} = X_{14} - X_7$

} Variables need to impose the restriction required to test the hypotheses.

Model 1 $$Y_1 = a_0U + a_1X_1 + a_8X_8 + a_9X_9 + a_{10}X_{10} + a_{11}X_{11} + a_{12}X_{12} + a_{13}X_{13} + a_{14}X_{14} + E$$

Restrictions: $$\left[(a_7 + a_9) - (a_8 + a_{10})\right]/2 = \left[(a_{11} + a_{13}) - (a_{12} + a_{14})\right]/2$$

Model 2 $$Y_1 = a_0U + a_{15}X_{15} + a_{16}X_{16} + a_{17}X_{17} + a_{18}X_{18} + a_{19}X_{19} + a_{20}X_{20} + a_{21}X_{21} + E$$

If the researcher finds a significant F-value and if the value of the left side of the restriction is greater than the value of the right side of the restriction, the researcher would conclude that the data supports the hypothesis. Without the knowledge of regression, the researcher may not even ask such a question, let alone be able to test it, even though the question may be of great importance to his study.

4. A fourth advantage of regression is that by using the multiple linear regression procedures, questions that involve covariance are easier to test and interpret (Kerlinger, 1973; Kerlinger and Pedhauzer, 1973; Ward and Jennings, 1973; Williams, 1974; Draper and Smith, 1966; Newman, et al., 1976; McNeil, 1976). This point can be demonstrated by the procedure listed below:

Example 3:

Hypothesis I: The posttest scores for the experimental group are significantly higher than the posttest scores for the control group over and above the differences due to I.Q. scores. (The variables listed in Example 1 are also used for this example)

Model 3

$$Y_1 = a_0U + a_1X_1 + a_2X_2 + a_3X_3 + E$$

Restriction: $a_1 = a_2$

Model 4

$$Y_1 = a_0U + a_3X_3 + E$$

If Model 3 is found to be significantly different from Model 4, this would indicate that there is a significant difference between the groups. Also, if $a_1 < a_2$, this would suggest that the Experimental Group had higher posttest scores than did the Control Group (at some specific α level).

5. Another advantage of regression is that it facilitates the calculation and interpretation of trends (functional relationships). When the research question of interest is one of trends or functional relationships, one often finds the use of inappropriate statistical models which cannot accurately reflect the research questions (Newman, 1974).

When researching developmental questions, one is often more interested in functional relationships than mean differences. There is generally a continuous variable that

is of interest, such as time, age, population sizes, I.Q. When traditional analysis of variance is employed, for example, continuous variables are forced into categorizations. This causes the researcher to lose degrees of freedom, and there is a potential loss of information. This loss is contingent upon how representative the categories are of the inflections in the naturally occuring continuous variable.

Since continuous variables are frequently artifically categorized, the analysis produced by such a procedure may not really reflect the researcher's question or interest. The most efficient method for writing statistical models that reflect trend or curve fitting questions, is the general case of the least squares solution, linear model (Multiple Linear Regression Procedures, Newman (1974), McNeil, Kelly, McNeil (1975), Draper & Smith (1966), Kelly, Newman, and McNeil (1973)). This procedure allows one to write linear models, which specifically reflect the research question.

Linear Regression is an excellent statistical tool for looking at a population trend or comparing multiple trends over time (Newman (1974), Erving (1975)).

For Example, in Figure 1, a graph is presented that reflects the researcher's interest in learning if there are significant differences in trends (in this case slope differences) between subjects who received a Developmental Reading Program (X_1) and students who did not receive the

Program (X_2), as it relates to their cumulative G.P.A.

Example 4:

Hypothesis I: There are significant differences in slopes for X_1 and X_2 in predicting the student's cumulative G.P.A.

The models needed to test this hypothesis are as follows:

Model 1 $Y_1 = a_0U + a_2X_2 + a_3X_3 + a_4X_4 + E$

Restriction: $a_3 = a_4$

Model 2 $Y_1 = a_0U + a_1X_1 + a_2X_2 + a_5X_5 + E$

Y_1 = cumulative G.P.A.

X_2 = 1 if student had program, 0 otherwise

X_2 = 1 if student did not have program, 0 otherwise

X_3 = number of the quarter hours for the subjects who had the program, 0 otherwise

X_4 = number of the quarter hours for the subjects who did not have the program, 0 otherwise

$X_5 = X_3 + X_4$ = number of quarter hours for all subjects

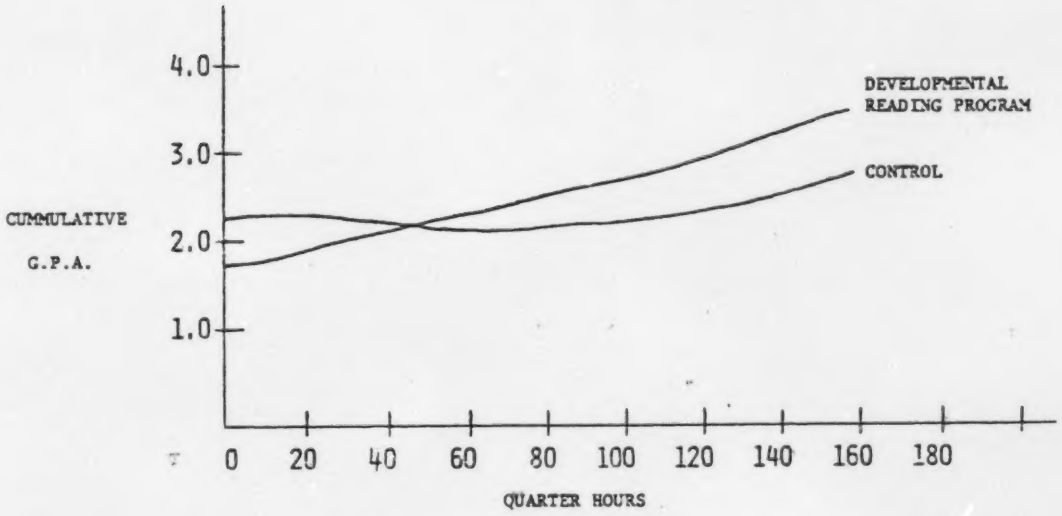
U = unit vector, 1 if subject us in the sample, 0 otherwise

$a_0, \ldots a_5$ = partial regression weight

$E_{1,2}$ = error $Y - \hat{Y}$

FIGURE 1

TREATMENTS AND CUMMULATIVE G.P.A. FOR STUDENTS



If Model 1 is found to be significantly different from Model 2, that is, the F-value is significant, this would indicate that there is a significant difference between students who took the program and students who did not take the program in terms of their cumulative G.P.A.

Regression will also allow many other questions to be asked when dealing with trend analysis. Second degree or third degree relationships (curvilinear relationships) could be investigated. Regression models could be written that would reflect such trends.

6. The applied statistician and researcher is plagued with the problem of disproportional cell sizes in factorial experimental designs. This may occur because of mortality in the laboratory animals being used in the experiment; the required number of subjects not available; someone who had agreed to take part in the experiment fails to show up; or the data may represent the proportionality that exists in the "real world." (Newman, Oravecz,1977)

When the researchers feel disproportionality is severe enough to be of concern, there are a variety of procedures that he can utilize to attempt to correct for the potential problems. However, before any corrections are applied, one should be sensitive to the underlying assumption that they are making about the population from which their data is drawn, and the investigator must also be very clear about the research question he is interested in asking.

The following is a check list adopted from Newman and Oravecz (1977), of the type of information that a researcher should investigate before selecting a method for correcting for disproportionality:

a. know something about the theoretical and/or empirical relationship between the variables being studied;

b. know some of the descriptive data about the population one wishes to generalize to in relation to the specific variables being studied;

c. know the specific research question under investigation if one decides an adjustment for disproportionality

is needed, then

d. know the underlying assumptions and implications for different adjustment procedures, and

e. know the consequences for using the selected adjustment procedure on the interpretation and generalization of the data.

A detailed discussion of the underlying assumptions can be found in the article by Newman and Oravecz (1977).

There are a variety of solutions to the unequal N's problem, which can be divided into two major categories--approximate and exact.

Examples of approximate solutions are: randomly eliminating data and running the analysis on just group means, therefore, decreasing the number and power. A researcher using any of these solutions is generally aware of the limitations and problems.

What may be more misleading are the exact solutions which are all technically correct but which, like the mean, median, and mode, are answering different questions. The three exact solutions, which are listed below, are the full rank method, fitting constants method, and the hierarchial method (Newman & Newman, 1975).

Example 5:

A. Solution 1 - Full Rank Solution

A symbolic example of this procedure is presented below for a two factorial design.

Model 1 $Y_{kab} = \delta + b_1\alpha_a + b_2\beta_b + b_3\alpha\beta_{ab} + \varepsilon_{kab}$

Model 2 $Y_{kab} = \delta + b_4\beta_b + b_5\alpha\beta_{ab} + \varepsilon_{kab}$

Model 3 $Y_{kab} = \delta + b_6\alpha_a + b_7\alpha\beta_{ab} + \varepsilon_{kab}$

Model 4 $Y_{kab} = \delta + b_8\alpha_a + b_9\beta_b + \varepsilon_{kab}$

Y_{kab} = is the score for subject k in row a and column b

δ = is the grand $\bar{X}$

α_a = is the effect for row "a"

β_b = is the effect for column "b"

$\alpha\beta_{ab}$ = is the interaction effect for the row "a" and column "b"

ε_{kab} = is the error term for each subject

$b, \ldots b_n$ are partial regression coefficients

Adjustment for Solution 1

Adjustment for A main effects test Model 1 against Model 2

Adjustment for B main effects test Model 1 against Model 3

Adjustment for A*B effects test Model 1 against Model 4

B. Solution 2 - Fitting Constants Method

The following is a symbolic representation of this solution:

Adjustment for Solution 2

Model 4 $Y_{kab} = \delta + b_{10}\alpha_a + b_{11}\beta_b + \varepsilon_{kab}$

Model 5 $Y_{kab} = \delta + b_{12}\beta_b + \varepsilon_{kab}$

Model 6 $Y_{kab} = \delta + b_{13}\alpha_a + \varepsilon_{kab}$

Adjustment for A main effects test Model 4 against Model 5

Adjustment for B main effects test Model 4 against Model 6

Adjustment for AB interaction effects test Model 4 against Model 1

C. Solution 3 - Hierarchial Method (Cohen (1968), Williams (1974).

The following is a symbolic representation of this solution:

Adjustment for Solution 3

Model 7 $Y_{kab} = \delta + b_{14}\alpha_a + \varepsilon_{kab}$

Model 8 $Y_{kab} = \delta + \varepsilon_{kab}$

Model 9 $Y_{kab} = \delta + b_{15}\alpha + b_{16}\beta + \varepsilon_{kab}$

Adjustment for A main effects test Model 7 against Model 8

Adjustment for B main effects test Model 9 against Model 7

Adjustment for AB interaction test Model 1 against Model 9

Each of the three least square solutions make different assumptions about the meaningfulness and "usefulness" of the correlations between the A main effect, B main effect, and AB interaction.

Solution 1, for example, when testing the A main effect, assumes the correlation between A and B and the AB interaction is of an accidental nature, and therefore should not be considered (Rock, et al., 1976). This solution is most likely to be prefered when one can assume that the missing subjects producing disproportionality were random. If one is unable to make this assumption then it would be inappropriate to use Solution 1, (which may be the case most frequently).

Solution 2 assumes that there is no correlation between the A and B main effects in the population. Therefore, the correlation between A and B in the sample is a function of disproportionality and not representative of the population. Solution 2 then attempts to adjust for this correlation.

However, Solution 2 assumes that the correlations

between the main effects and the interaction, which results from the disproportionality, are not spurious and are characteristic of the population. Therefore, it does not attempt to adjust for this correlation.

If one cannot assume that the correlations between the A and B main effects due to disproportionality are due to chance, then Solution 2 would be an inappropriate correction.

Solution 3 requires an a priori ordering of the importance of each variable. Let us assume that the a priori ordering are: A main effects, B main effects, AB interaction, respectively (Newman and Oravecz, 1977).

It is important to determine which of these methods are reflecting the question that we are interested in answering. One can only do this by being sensitive to one's research question and by being aware of the different statistical techniques which are more appropriate than others.

Methods That Can Be Utilized To Deal With The Concerns Of Correlation, Upward Bias R^2 Values, And Multicollinearity

There are three concerns which have been expressed by the critics of multiple linear regression that have drawn a great deal of attention. One concern expressed by some critics is that causality cannot be inferred from studies that use regression procedures. Another concern that has been expressed by some researcher is the

tendency for multiple correlations to be upward bias. The third major concern, called multicollinearity, problems produced by the non-orthogonality of the independent variables. [Note: One of the problems with disproportionality, a concept discussed in the preceding section of this paper, is that disproportionality produces correlation between its variables, i.e., multicollinearity.]

It is the purpose of this section of the paper to present a discussion of possible methods that a researcher could use in order to deal with these problems.

1. One of the concerns that has been expressed by the critics of multiple linear regression is that one cannot infer causation if regression or correlation is used. This concern which has been expressed both formally and informally, can be found in a recent article entitled "Regression Analyses and Education Production Functions: Can They Be Trusted?" The authors Lyecke and McGinn (1975) conclude that a researcher cannot appropriately infer causation from regression techniques.

The statement by Lyecke and McGinn (1975) is correct. However, causation cannot be inferred from any statistical tool unless an appropriate research design is utilized. If causation is to be inferred, regression as a statistical tool, as is the case for any other statistical tool, must be used in relationship with some research design that can be found, for example, in Stanley and Campbell (1969).

To the extent that this design has internal validity, the researcher can infer causal relationships between the independent variables and dependent variables.

If a research design is ex post facto, where the independent variable is not under the control of the researchers, no matter what technique is used, one cannot infer causation. It is not technically legitimate to infer causation when the design is ex post facto. Even though a variety of statistical techniques such as path analysis as developed by Blalock (1962, 1964, 1970, and 1972) and more recently component analysis developed by Mood (1971), have attempted to get at causal relationships of ex post facto data, through the manipulation of regression techniques, one still cannot technically infer causation (Newman & Newman, 1975). Newman and Newman stated the following with regard to causation and component analysis:

> Since one of the major purposes for calculating component analysis is to attempt to improve the explanation of ex post facto research designs, this can lead one to mistakenly believe that the unique variance accounted for by an independent variable with a criterion is of a causal nature (p. 45).

In a similar fashion, Lee Wolfle (1977), states the inability of a researcher who is using path analysis on ex post facto data to infer causality as follows:

> Although path analysis is a method for considering cause, neither it, nor any other method, can be used for inferring causality from non-experimental data (p. 39)

It is, therefore, not the use of multiple linear regression that precludes the researcher from infering causal relationships between the variables. It is the lack of a true experimental design that prevents the researcher from making such inferences. Causation can only be inferred if a true experimental design was utilized, irregardles of the statistical tools that were used to analyse the data.

2. Many researchers mistakenly believe it is meaningful to include in their reports only that a manipulation of an independent variable was shown to have a significant effect upon a dependent variable. The magnitude of this effect is not given to the reader. The magnitude of this effect, which could be presented by citing the R^2 or η^2 values, must be taken into consideration when a researcher is interpreting the practical significance of experiment results (Byrne, 1974, Cohen, 1969, Fredman, 1972).

Most researchers are aware that a R^2-value tends to be higher in the sample than in the population from which the sample was drawn. This shrinkage is due to the fact that the regression weights are calculated to maximize the prediction of the criterion. The sampling error is capitalized on when calculating the regression weights, so that the predictive power for any one sample is maximized. It should be noted, however, that in an article by Dalton (1977) it was suggested that this

overestimation of R^2 may not be too great in many cases and is really a good estimate of the population value.

Dalton (1977) used Monte Carlo methods to compare $\hat{\omega}^2$, $\hat{R}^2$ (R^2 after a shrinkage formula has been applied), and R^2. The bias in R^2 was consistently positive and it decreased as the sample size increased. However, Dalton concluded from his study that even though $\hat{R}^2$ and $\hat{\omega}^2$ were superior to R^2 when $n \leq 30$, R^2 showed little bias in large samples. Therefore, this study may suggest that the upward bias tendency for R^2 values is not as prominent a problem as once thought. However, there was one shortcomming of Daltion's (1977) study and that is he only examined at the three variable situation. This greatly limits the possible generalizablity of the study.

There are four possible methods which can be used to obtain a corrected R^2 ($\hat{R}^2$). These methods are entitled the Wherry Method, McNamara Method, Lord Method, and Cross-Validation Method. Uhl and Eisenberg (1970) empirically investigated the accuracy of three of these methods; Wherry's original formula (1931), McNemar's modification (1962), and Lord's (1950) formula. These formulas are:

$$\hat{R}^2 = 1 - (1 - R^2)\frac{N-1}{N-K} \qquad \text{(Wherry)}$$

$$\hat{R}^2 = 1 - (1 - R^2)\frac{N-1}{N-K-1} \qquad \text{(McNemar)}$$

$$\hat{R}^2 = 1 = (1 - R^2)\frac{N+K+1}{N-K-1} \qquad \text{(Lord)}$$

where: $\hat{R}$ = the corrected estimate of the multiple correlation

R = the actual calculated multiple correlation

K = the number of independent variables

N = the number of independent observations

Uhl and Eisenberg found that even though Wherry's and McNemar's formulas are the most commonly used, Lord's formula consistently gave more accurate estimates for the five different N sizes they investigated (N = 50, 100, 150, 250, 325) and for the situations using two through thirteen predictor variables.

A study conducted by Klein and Newman (1974) indicated that when there are 100 subjects for each variable all three formuli produce the same estimates.[1] When the ratio is less than that, Lord's formula is consistently more conservative, that is, it shrinks more. As the variables increase, there seems to be a tendency for McNemar and Wherry to produce more similar results.

---

[1]In a discussion between Keith McNeil and Isadore Newman, the topic of the ratio between variables and subjects was reviewed. McNeil stated that this ratio may not be equivalent for continuous variables and dichotomous variables. McNeil suggested that in order to establish a 10:1 ratio for a continuous variable, one may have to have ten subjects per "grouping" of the variable, that is, if the values of a continuous variable distribute themselves into approximately three distinct groups, ten subjects are needed for each group in order to retain a 10:1 ratio. This is probably a conservative estimate, but there is no data to empirically support the claim of equivalent ratio for continuous and dichotomous variables.

Klein and Newman further stated that it is conceptually meaningless to interpret negative R^2, and since the lowest possible R^2 one can legitimately obtain is 0, it seems that these formuli need a correction factor added so that they are bounded on the low and by $R^2 = 0.0$ and on the high end by $R^2 = 1.0$. It is therefore suggested that if one uses any of these three shrinkage estimates that any negative R^2 be interpreted as if it were $R^2 = 0$.

Kelly, et al. (1969), suggest cross validation procedures as estimates of shrinkage instead of using the more mathematical approaches used by Wherry, McNemar, and Lord. The cross-validation procedure estimates the shrinkage by applying the weighting coefficients from the original sample to a new sample of subjects from the same population.

For example, assume the weights for Model 1 in Example 3 are as follows:

Example 6:

Model 1 $\quad Y_1 = 10U + 6.85\ X_1 + 5.00X_2 + .05\ X_3 + E$

A new sample should be taken from the same population and the variable X_6 (the predicted criterion) should be generated for this sample by using the weights obtained from the first sample. The transformation needed would be as follows:

$$X_6 = 6.85 * X_1 + 5.00 * X_2 + .05 * X_3 + 10$$

If the correlation between X_6 (the predicted criterion) and Y_1 (the observed criterion) was as high as the R-value for Model 1, the researcher could consider the R^2-value for Model 1 to be stable.

Some of the differences in the shrinkage estimates, using the different procedures amy be explainable. For example, Wherry's and McNemar's formulas both attempt to estimate the population R, based on the sample, while Lord's formula attempts to estimate the R from the sample to antoher sample. This is conceptually similar to the Cross Validation procedure suggested by Kelly. In deciding which method of estimating shrinkage is to be used, it is important to consider the underlying assumptions of each procedure (Klein and Newman, 1973). That is, cross validation will tend to be more conservative estimation. Thus, it will tend to produce larger shrinkage in R^2. If one is interested in making predictions based on one sample to another sample, cross validation and Lord's approach tend to be the better estimates. However, if one wants to estimate population values from a sample, Wherry's and McNemar's approaches would be preferable.

3. The third major concern, which has drawn a great deal of attention, is multicollinearity, that is, a situation in which the predictor variables are nonorthogonal. One of the problems that multicollinearity can cause is large standard errors in the sampling distributions of the standardized regression coefficients. These large standard errors allow small changes in the relationships between independent variables from sample to sample to produce large regression weight differences even though their signs tend to be stable. Therefore, interpreting regression weights can be highly misleading due to this high variability (McNeil, et al., 1975). Another problem caused by multicollinearity is that a researcher is more likely to committee a Type II error. (Vasen & Elmore, 1975).

There are a number of ways to deal with the problem of multicollinearity. Five such methods are:

a. component regression
b. factor regression
c. ridge regression
d. "benign neglect"
e. a system of equations.

III. A. One method suggested in the literature for dealing with multicollinearity is component analysis (Newman and Newman, 1975; Massy, 1965). Component analysis is a procedure which divides variance into two proportions; Unique variance(Uq) is the proportion of variance attributed to a particular variable when entered last into the regression equation. Common variance (Cv) may be

conceptually thought of as the degree of overlap of correlated variables in the prediction of the criterion. Any given common variance must be independent of unique and other common variance.

The calculation of unique variance for three predictor variables could be handled as follows:

Example 7:

Let

Y_1 = grade point average

X_1 = SAT score

X_2 = I.Q. score

X_3 = high school class work

The number of independent components can be calculated by the equation:

$$2^N - 1$$

where: N = the number of predictor variables

Thus, for this example, the number of independent components would be equal to:

$$2^3 - 1 = 7$$

The number of sets of unique variance is equal to the number of predictive variables. For this example, there would be three sets of unique variance [Uq (1), Uq (2), Uq (3)]. The number of second, third, etc., order variance can be determined by the following formula:

$$NC_n = \frac{N!}{n!(N-n)!}$$

where: N = number of predictor variables

n = number of variables taken at a time

NC_n = number of combinations of N objects, taking n number at a time, independent of order.

In this example, the number of second and third order commonalities are equal to the following:

$$NC_n = \frac{3!}{2!\ (3-2)!} = 3 \qquad NC_n = \frac{3!}{3!\ (3-3)!} = 1$$

The three sets of second order commonality are $Cv_{(1,2,)}$, $Cv_{(1,3)}$, $Cv_{(2,3,)}$; and the third order commonality variance is $Cv_{(1,2,3)}$.

These components are additive and when summed will equal the total proportion of variance accounted for by the R_F^2 of the full model. Mood (1969) developed a rule for determining the R_s^2 necessary for caluclating unique and common components of variance. The rule is to develop products of the variables being considered.

For example, if one is interested in the $Uq(X_1)$ in this example with three predictor variables (X_1, X_2, X_3), first subtract that variable of interest (X_1) from one, multiplied by a -1, and multiple other variables in the equation.

rule: $-1(1-X_1)\ X_2, X_3 =$

$$-X_2X_3 + X_1X_2X_3$$

Next, take the variables that are a product of the expansion and calculate the R_s^2 that is indicated by each

set (separated by + and - signs)

$$Uq(X_1) = -R^2_{y\cdot 23} + R^2_{y\cdot 123}$$

In a similar manner, one of the second order and the third order commonality variances would be calculated as follows:

rule: $-1(1-X_1)\ (1-X_2)\ X_3 =$

$$-X_3 + X_1X_3 + X_2X_3 - X_1X_2X_3$$

$$Cv_{(1,2)} = -R^2_{y.3} + R^2_{y.13} + R^2_{y.23} - R^2_{y.123}$$

rule: $-1(1-X_1)\ (1-X_2)\ (1-X_3) =$

$$-1 + X_1 + X_2 - X_1X_2 + X_3 - X_1X_3 - X_2X_3 + X_1X_2X_3$$

$$Cv_{(1,2,3)} = R^2_{y.1} + R^2_{y.12} - R^2_{y.3} - R^2_{y.23} + R^2_{y.123}$$

[Note: When a <u>one</u> is by itself in the expansion, it is ignored in determining which R^2_s should be calculated.]

For further details in how to calculate component analysis, see Mood (1969, 1971), Kerlinger (1973) and Houston and Bolding (1975).

With all techniques, one must be aware of the limitations so that the technique can be employed most efficiently. The following are some of the limitations one should be sensitive to when using component analysis (Newman and Newman (1975)):

1. As in the example, when there are three predictor variables, there will be seven components. One can easily see the rather large number of R^2s that have to be calcu-

lated for just three predictor variables in the full model. However, in using multiple regression, the investigator frequently has many more than three predictor variables. Therefore, the number of components can easily become impractical to handle.

2. An integral part of component analysis is the concept of Uq. Uq is operationally defined as:

> variance accounted for by a variable when entered last in a multiple regression equation.

Therefore, the Uq depends upon and is affected by the variables that are already under investigation. Even though the Uq is independent, in the set of variables for that sample, the variable is not independent.

3. As the number of predictor variables increase, the number of higher order commonality components also increase. Just as it is difficult to interpret higher than third order interactions in traditional analysis of variance, it is also difficult to interpret higher than third order commonalities.

4. In examining some of the formuli for calculating the commonality components, one becomes sensitive to the possibility that some of the components can easily account for a negative proportion of variance. When this

situation is encountered, it becomes very difficult to interpret or make conceptual sense out of the analysis.

5. Mood (1971) stated an important limitation one should consider. The unique variance (Uq) accounted for by an independent variable can change radically from situation to situation. However, the Uq attributed to a factor that the variable is a part of is not likely to change. Therefore, Mood suggests that one should group the variables based on the underlying concept they seem to be measuring. This would produce a more stable estimate. This group process will also have a side benefit of reducing the total number of predictor variables which will make the component analysis much more manageable. However, if one uses the procedure suggested by Mood, the weighting of each variable becomes a problem. Do the factors account for the same 100 percent of the proportion of variance accounted for when each variable is used separately? If not, one is loosing possibly significant information. Finally, it is difficult to decide on which variables should go together. Quite often, variables that look as if they are measuring the same underlying construct, are not.

III. B. Another method by which a researcher can deal with the problems caused by nonorthogonal predictor variables is factor regression.

Factor multiple regression is a procedure that may circumvent some of the problems associated with component regression (Massy (1965), Duff, Houston and Bloom (1971),

Connett, Houston and Shaw (1972), Newman (1972)). It is a method that enables one to empirically determine the factors with which the variables are associated.

The first step in the procedure is to orthogonally factor a set of independent variables into a nXn factor matrix. Connett, et al. (1972) suggests that this factor matrix may be rotated, but only with a rotation that preserves the orthogonality of the factors. The next step is to standardize the independent variables. This matrix of standardized variables is postmultiplied by the matrix to obtain the factor variables. Because these factor variables are orthogonal, the beta weights of these variables, when used in a regression equation, will tend to be stable. Therefore, this procedure allows greater interpretation of the beta weights to be made.

An additional advantage of using factor scores is that when a matrix is factored much of the error variance tends to be distributed in the factors that account for the least variance. Therefore, one of the possible by-products of using factor scores which account for most trace variance as predictors is the likelihood of increasing reliability; therefore, decreasing shrinkage (Newman, 1972).

If one is interested in improving the multiple regression equation by using factor techniques, there is only one way this can be done. That is, the number of factors used must be less than the number of original

variables. This will increase the df and also possibly decrease shrinkage-estimates. Because of this, some researchers have used only the few factors that account for the "greatest" amount of the factored trace. However, when this is done, one may be losing information that can account for criterion variance by eliminating a factor that accounts for very little trace of the factored matrix but is highly correlated with the criterion scores.

Using only the factors that account for most of the trace should be avoided when the prdictor variables that are being factored are likely to be highly reliable. Some examples of such variables are: height, weight, religion, sex, income, age, etc. Under these conditions, a variable that accounts for little of the trace variance may be a good and highly reliable predictor of criterion variance.

When using factor regression one should be aware of when it can be most appropriately used. It is the authors' opinion that the factor regression approach may be more appropriate than component analysis when one is interested in determining the unique variance accounted for, especially when the number of predictor variables is relatively large and there are a minimum of ten subjects for every variable. However, if one is interested in the commonality, the factor regression procedure is not appropriate. In this case, if one has a large number of

variables and subjects, it is possible to use factor analysis with oblique rotation. This procedure will condense the large number of variables into factors which can be used as a new set of predictor variables. Since these factors may be oblique (correlated), one may then wish to do a component analysis which will yield estimates of the unique and common variance attributed to the factors. Obviously, the oblique solutions lack many of the desirable characteristics which make the orthogonal solution easier to interpret. However, there are times when a researcher may be interested in the common proportion of variance attributed to factors which are theoretically and empirically related.

III. C. A method called ridge regression has been proposed as a possible means by which a researcher can obtain stable regression coefficients (Hoerl (1962), Hoerl and Kennard (1970 (a), 1970 (b)), Marquardt and Snee (1975)). The ridge regression procedure requires that a constant be repeatedly added to the diagonal of the $\overline{X}^{1}X$ matrix (where the X variables are scaled so that $\overline{X}^{1}X$ has the form of a correlation matrix) before the matrix is inverted. That is, consider the standard model for multiple linear regression

$Y = X\beta + E$

X = nXp matrix of p predictor variables at each n data points

Y = vector of observed values

β = pX1 vector of population values of the parameters

ε = nx1 vector of experimental errors (E (ε) = 0)

where

$$\hat{\beta} = (X'X)^{-1}(X'Y)$$

$X'X$ = the product of transposed X and X

$X'Y$ = the product of transposed X and Y

$\hat{\beta}$ = least squares estimator of β

Ridge regression, as described in more detail in Hoerl (1962) and Hoerl and Kennard (1970a, 1970b) is an estimation procedure based upon

$$\hat{\beta}^* = (X'X + KI)^{-1}(X'Y)$$

I = identity matrix

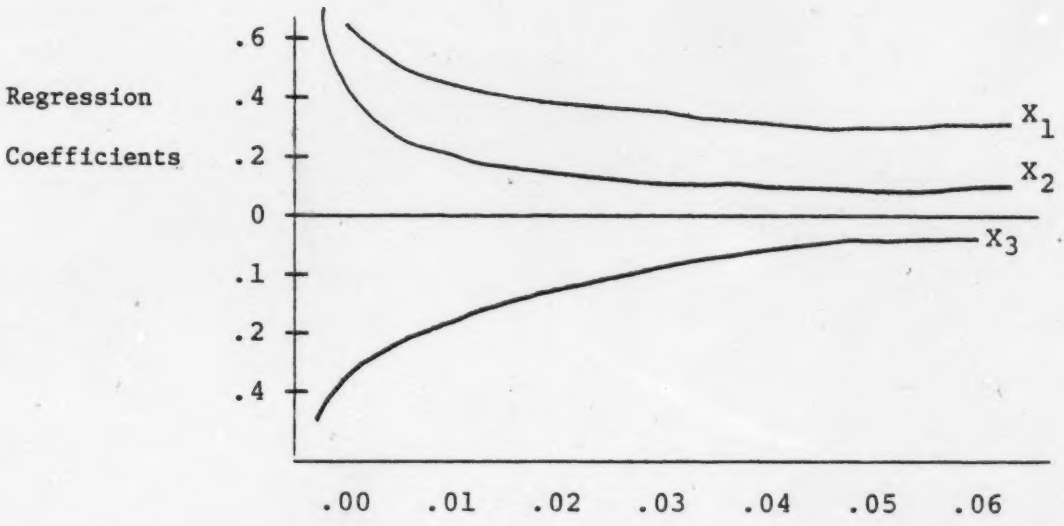
$K = 0 \leq K < 1$

$\hat{\beta}^*$ = ridge estimator of β

where K is a conststant number added to the identity matrix I. The researcher can determine the appropriate K value, i.e., the K value that stabilizes the regression coefficients by examining the Ridge Trace. The Ridge Trace is a plot of the coefficient weights vs. the K values. A hypothetical diagram of the Ridge Trace is given in Figure 2 for the variables X_1, X_2, and X_3.

FIGURE 2

RIDGE TRACE



At the K-value where the ridge traces for the variables appear to become approximately parallel the regression coefficients become stable. In Figure 2 the ridge traces become approximately parallel where K = .04. Thus, the researcher would use the regression coefficients that correspond to that point.

The researcher will find that for models with low R^2 values require larger values of K than do models with high R^2 values. Also, increasing K indefinitely will ultimately force all coefficients to zero, but it is not uncommon to see a coefficient (usually after an initial sign change) to increase in absolute value as K increases (Marquardt and Snee, 1975).

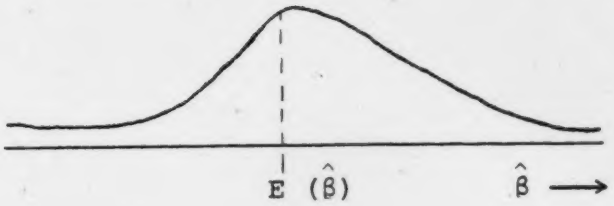
Before this procedure is used, however, a researcher should be aware of the differences between the coefficients produced by the least squares solution and coefficients produced by ridge regression. The least squares solution yields coefficients that minimize the residual sum of squares. The expected value of the coefficients are unbias ($E(\hat{\beta}) = \beta$) and have the minimum variance among all linear unbiased estimators (see Figure 3a).

In ridge regression, the variance of the coefficients decreases (see Figure 3b) as the value for K increases. However, the bias of this estimator increases ($E(\hat{\beta})$ β) as the value of K increases. What the researcher is doing with ridge regression is accepting a little bias in the expected value of the coefficient in return for a lower mean square error [MSE = variance of the coefficient + $(\text{bias})^2$]. In fact, the objective of ridge regression is to find a value of K which gives a set of coefficients with smaller MSE than the one produced by the least squares solution. As the K value increases, the residual sum of squares will increase. But remember, it is not the objective of ridge regression to obtain the "best fit" for the sample data but rather to develop stable coefficients (Marquardt and Snee, 1975).

FIGURE 3a

(Marquardt and Snee, 1975)

VARIANCE AND BIAS IN AN ESTIMATOR

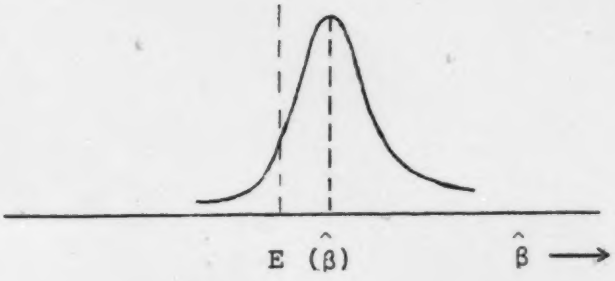


Zero Bias

Large Variance

$E(\hat{\beta})$

FIGURE 3b



Non-zero Bias

Small Variance

$E(\hat{\beta}) = \beta$

The preceding discussion on the differences between the least squares solution and ridge regression does point out one limitation to using ridge regression. Because the espected values of the coefficient is bias, hypothesis testing would be questionable. Thus, if the researcher is attempting to test hypotheses, ridge regression may not

be the correct method for him to use in handle the problem of of multicollinearity (unstable coefficients).

III. D. A fourth possible method for dealing with the problem of multicollinearity is <u>not</u> to deal with it! The argument for this position can be demonstrated by the following example:

Example 8:

Y_1 = posttest score

X_1 = I.Q. score

$X_1^2 = X_1 * X_1$

$E_{1,2,3}$ = error for each subject

U = unit vector

$a_o . . . a_2$ = regression coefficient weights

<u>Model 1</u> $Y_1 = a_oU + a_1X_1 + a_2X_1^2 + E$ $R_1^2 = 1$

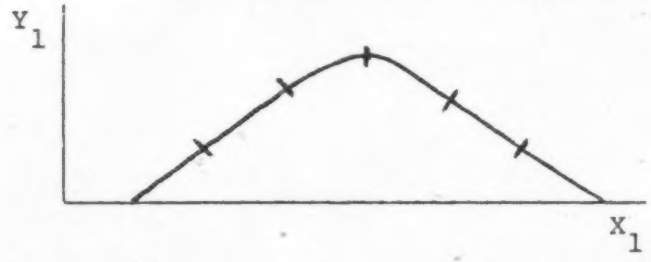
<u>Model 2</u> $Y_1 = a_oU + a_1X_1 + E$ $R_2^2 = 0$

<u>Model 3</u> $Y_1 = a_oU + a_2X_1^2 + E$ $R_3^2 = 0$

Assume that the variable X_1 is related to variable Y_1 in the manner indicated in Figure 4.

FIGURE 4*

THE RELATIONSHIP BETWEN X_1 AND Y_1



*Note: Example in Figure 4 was given by John Pohlman at the 1977 A.E.R.A. Convention to support the argument that forward Stepwise Regression may be eliminating surpressor variables.

The R_1^2 value for Model 1 would be equal to one. However, the R_2^2 value for Model 2 would be equal to zero for the relationship between X_1 and Y_1. Also, Model 3 would have an R^2 value equal to zero for the relationship between X_1^2 and Y_1.

Model 1 is attempting to account for the variance in Y_1 by using variables X_1 and X_1^2. It is important to note that the correlation between X_1 and X_1^2 is high. That is, multicollinearity is present in Model 1. In Model 2 and Model 3, the multicollinearity is eliminated by the traditional procedure of eliminating one of the correlated independent variables. What has also been eliminated, however, in both Model 2 and Model 3 is a surpressor variable (Surpressor variables have also been called in intervening variable or a sleeper variable).

A surpressor variable is present when a variable has a low correlation with the criterion and is highly correlated with some other variable in the predictive equation. In addition, when this variable is placed in the predictive equation along with the variable with which it is highly correlated, the R^2 of the predictive equation will increase significantly. Such a surpressor variable is present in Model 1. When both X_1 and X_1^2, variables which are highly correlated, are used together as they are in Model 1, the R^2-value of Model 1 increases significantly over the R^2 values of Model 2 and Model 3. The point is

that the researcher does not want to eliminate a surpressor variable.

Consider the following example:

Example 8:

Let Y_1 = achievement scores

X_1 = treatment group

X_2 = control group

X_3 = reaction time

Assume that X_1 is correlated with X_3 and X_2 is also correlated with X_3.

The hypothesis to be tested is as follows:

H_1: There is a significant difference between the achievement scores for the control group and the experimental group over and above the differences due to reaction time scores.

The models below:

the hypothesis:

Model 1

$$Y_1 = a_oU + a_1X_1 + a_2X_2 + a_3X_3 + E$$

Restrictions: $a_1 = a_2$

Model 2

$$Y_1 = a_oU + a_3X_3 + E$$

It is important to note, however, that the researcher must select his variables carefully. That is, his hypothesis should probably not include X_3 if the relationships between X_3 and the groups (X_1 and X_2) are not found in other research in the discipline or are illogical. If these relationships are not usually found or are unstable, the results of the hypotheses tests may vary from sample to sample.

The researcher should also be aware that he is more likely to commit a Type II error when the relationship between X_3 and the groups is not consistent across the continuum of Y. That is, there is an interaction between groups and the reaction time. In fact, when there is an interaction between groups and reaction time, one of the conditions of covariance has been violated (homogenuity of regression) and, therefore, analysis of covariance is no longer appropriate.

III. E. Soper (1976) suggested in a review of a study on the use of programmed instruction in economics that a system of equations should be established in order to correct for the nonorthogonality of the independent variables. For example, consider the following hypothesis and variables:

Example 9:

H_1: There is a significant difference between the control group and the experimental group in posttest scores over and above the difference due to scholastic ability.

Y_1 = posttest score

X_1 = pretest score

X_2 = SAT score (Scholastic Aptitude Test score)

X_3 = experimental group

X_4 = control group

a_o. . .a = regression coefficient weights

E_1. . .E = the error terms (Y - Y) for the different models

The traditional method of analyzing the data would be to test Model 1 against Model 2.

Model 1

$$Y_1 = a_oU + a_1X_1 + a_2X_2 + a_3X_3 + a_4X_4 + E_1$$

$$a_3 = a_4$$

Model 2

$$X_1 = a_oU + a_1X_1 + a_2X_2 + E_2$$

If X_3 and X_4 are correlated with X_1 (SAT scores), Soper (1976) would suggest that a system of equations, or in this case on equation, would need to be specified. The needed equation would be as follows:

Model 3

$$X_2 = a_oU + a_3X_3 + a_4X_4 + E_3$$

The value for E_3 would represent the amount of variation in SAT scores that are unrelated to group membership.

Next, E_3 would be used as an independent variables in Model 4, and Model 4 would be tested against Model 5. Model 4 and Model 5 are as follows:

Model 4

$$Y_1 = a_oU + a_1X_1 + a_3X_3 + a_4X_4 + a_5E_3 + E_4$$

Restriction: $a_3 = a_4$

Model 5

$$Y_1 = a_oU + a_1X_1 + a_5E_3 + E_5$$

However, the researcher has not tested his original question of interest (Type VI error) which was: Is there a significant difference between the control group and the experimental group on posttest scores over and above the differences due to scholastic ability? What he has, in fact, tested is the hypothesis: There is a significant difference between the control group and the experimental group on posttest scores over and above the differences in I.Q. and SAT scores unrelated to group membership. This is a different question!

Also, one must be aware that correcting numerous variables for multicollinearity tends to make the interpretation of the results very difficult. The researcher may not be able to practically explain what a significant F-value indicates.

For example, assume two variables, I.Q. and sex, were correlated and the researcher set up a system of equations which included the following equation:

$$I.Q. = a_o U = a_1 \text{ sex} + E_6$$

The question is, How do we interpret E_6? It is whatever I.Q. is after sex has been removed from it. Is it still I.Q.? Most probably not.

In conclusion, the authors hope that this presentation which dealt with some of the currently identified problems in conducting research, speciffically when using regression, has sensitized the applied researcher to these problems and alternative solutions. The authors feel that no one paper can do justice to all the topics covered. However, we feel that this paper can be used as a guide to where one may go for more detailed information.

It should be kept in mind that the authors felt the regression approach is probably the most flexible and useful single tool available to the researcher. However, like any other tool, it is only as good as the insights and sensitivity of the user. Do not commit the classical error of asking a research question and testing it with a statistical model that is incapable of reflecting that question (Newman, et al., 1967).

REFERENCES

Blalock, H. M. Four-variable causal models and partial correlation. American Journal of Psychology, 1962, 68, 182-194.

Blalock, H. M. Causal inferences in non-experimental research. Chapel Hill: University of North Carolina Press, 1964.

Blalock, H. M. Path coefficients versus regression coefficients. American Journal of Psychology, 1967, 72, 675-676.

Blalock, H. M. (Ed.). Causal models in the social sciences. Chicago: Aldine, 1971.

Byrne, J. The use of regression equations to demonstrate causality. Multiple Linear Regression Viewpoints, 1974, 5(1), 11-22.

Campbell, D. & Stanley, J. Experimental and quasi-experimental designs for research. New-York: Rand McNally, 1969.

Cohen, J. and Chohen, P. Multiple regression as a general data analytic system. Psychological Bulletin, 1968, 70, 426-443.

Cohen, J. Statistical methods in research and production. New York: Hafner Publishing Co., 1961.

Connett, W., Houston, S. and Shaw, D. The use of factor regression in data analysis. Multiple Linear Regression Viewpoints, 1972, 2(4), 46-49.

Dalton, S. Shrinkage in R^2 and unbiased estimates of treatment effects using $\hat{\omega}$. Multiple Linear Regression Viewpoints. 1977, 7(3), 52-59.

Draper, N. R. and Smith, H. Applied regression analysis. New York: John Wiley and Sons, 1966.

Duff, W., Houston, S. and Bloom, S. A regression/principle components analysis of school outputs. Multiple Linear Regression Viewpoints. 1971, 1( ), 5-18.

Ervin, L. A multivariable approach to evaluation of a special educational program in higher education. Unpublished Dissertation, The University of Akron, Akron, Ohio, 1975.

Fraas, J. A proposed study of an economic survey course that utilizes simulations and simulation games. An unpublished research paper, Ashland College, Ashland, Ohio, 1977.

Friedman, H. Introduction to statistics. New York: Random House, 1972.

Hoerl, A. Application of ridge analysis to regression problems. Chemical Engineering Progress, 1962, 58(3), 54-59.

Hoerl, A. and Kennard, R. Ridge regression: biased estimation for nonorthogonal problems. Technometrics, 1970a, 12(1), 55-67.

Hoerl, A. and Kennard R. Ridge regression: applications to nonorthogonal problems. Technometrics, 1970b, 12(1), 69-82.

Houston, S. and Bolding, J. Part, partial and Multiple correlation in commonality analysis of multiple linear regression models. Multiple Linear Regression Viewpoints, 1975, S( ), 36-40.

Kelly, F., Newman, I. and McNeil, K. Suggested inferential statistical models for research in behavior modification. The Journal of Experimental Education, 1973, 41(4), 54-63.

Kerlinger, F. Foundations of behavioral research. (2nd ed.) New York: Holt, Rinehart and Winston, 1973.

Kerlinger, F. and Pedhazur, E. Multiple regression in behavioral research. New York: Holt, Rinehart and Winston, 1973.

Klein, M. and Newman, I. Estimated parameters of three shrinkage estimate formuli. Multiple Linear Regression Viewpoints, 1974, 4(4), 7-11.

Lord, F. Efficiency of sexiction when a regression equation from one sample is used in a new sample. Research Bulletin, 50, 40. Princeton, N.J.: Educational Testing Services, 1950.

Lyecke, D. and McGinn, N. Regression analysis and educational production functions: can they be trusted? Harvard Educational Review, 1975, 45 (3).

McNeil, K., Kelly, F. and McNeil, J. Testing research hypotheses using multiple linear regression. Carbondale, Illinois: Southern Illinois University Press, 1975.

McNemar, Q. Psychological statistics (3rd ed.). New York: Wiley and Sons, 1962.

Marquardt, W. and Snee, R. Ridge regression in practice. The American Statistician, 1975, 29(1), 3-20.

Massy, W. Principle components regression in exploratory statistical research. American Statistical Association Journal, 1965, 60, 23y-25b.

Mood, A. Macro-analysis of the American educational system. Operations Research, 1969, 17, 770-784.

Mood, A. Partitioning variance in multiple regression analyses as a tool for developing learning models. American Educational Research Journal, 1971, 8, 191-202.

Newman, I. A demonstration of multiple regression models that will facilitate the investigation of trends and functional relationships. A paper presented to the Psychology Division at the Ohio Academy of Science, April, 1974.

Newman, I. Some further considerations of using factor regression analysis. Multiple Linear Regression Viewpoints, 1972, 3(2), 39-41.

Newman, I. and Newman, C. A discussion of component analysis: its intended purpose, strengths and weaknesses. A paper presented to the American Educational Research Association, 1975; Multiple Linear Regression Special Interest Group.

Newman, I. and Newman, C. 38-22-36 Conceptual statistics for beginners. (3rd ed.) Akron, Ohio: The University of Akron, 1976.

Newman, I. and Oravecz, M. Solutions to the problem of disproportionality: a discussion of the models. A paper presented at the A.E.R.A., New York, April, 1977.

Newman, I., Deitchman, R., Burkholder, J., Sanders, R. and Ervin, L. Type IV error: inconsistency between the statistical procedure and the research question. Multiple Linear Regression Viewpoints, 1976, 6(4), 1-19.

Overall, J. and Spiegel, D. Concerning least squares analysis of experimental data. Psychological Bulletin, 1969, 72, 311-322.

Rock, D., Werts, C. and Linn, R. Structural equations as an aid in the interpretation of the nonorthogonal analysis of variance. Multivariate Behavioral Research, 1976, 11, 443-448.

Scheffe', H. The analysis of varicance. New York: John Wiley and Sons, 1959.

Soper, J. Second generation research in economics education: problems of specification and interdependence. The Journal of Economic Education, 1976, 7(2), 39-47.

Uhl, N. and Eisenberg, T. Predicting shrinkage in the multiple correlation coefficient. Educational and Psychological Measurement, 1970, 30, 487-489.

Vasu, E. and Elmore, P. The effect of multicollinearity and the violation of the assumption of normality on the testing of hypothesis in regression analysis. Multiple Linear Regression Viewpoints, 1975, 6(1), 21-45.

Ward, J. and Jennings, E. Introduction to linear models. Englewood Cliffs, New Jersey: Prentice-Hall, 1973.

Williams, J. Four-way disproportionate hierarchial models. Multiple Linear Regression Viewpoints, 1974, 5(2).

Williams, J. Regression analysis in educational research. New York: MSS Information Corporation, 1974.

Winter, B. Statistical principles in experimental design. New York: McGraw-Hill, 1962.

Wolfe, L. Path analysis and causal models as regression techniques: a comment. Multiple Linear Regression Viewpoints, 1977, 7(2), 33-40.